Emerging Microbes and Infections (2012) 1, e45; doi:10.1038/emi.2012.41

www.nature.com/emi

NEWS AND VIEWS

Pearls and pitfalls of genomics-based microbiome analysis

Nossa Carlos[1], Yi-Wei Tang[2] and Zhiheng Pei[3]

Emerging Microbes and Infections (2012) **1**, e45; doi:10.1038/emi.2012.41; published online 5 December 2012

Next generation sequencing (NGS) has supplanted traditional methods for microbiome analyses. Advantages include higher resolution and lower costs. Analyses have evolved from basic 16S rRNA surveys to metagenomic shotgun sequencing analyses. Limitations and obstacles still exist, e.g. 16S bias, bioinformatics bottlenecks, classification accuracies, and genome assembly using short reads. New tools are being developed to improve output and deal with larger amounts of data generated.

Ever since the germ theory of disease was validated by Koch in the late nineteenth century, our understanding of microbial diseases has been largely governed by the one pathogen/one disease paradigm until very recently when technical advances made it possible to scrutinize the entire microbial community (or microbiome) on all external (skin) and internal (mucosal) surfaces of our body.

Of the trillions of microbes inhabiting the human body, some are beneficial to us, some are neutral and others are harmful. Normally, they maintain a balanced community structure and a symbiotic relationship with the host, but alteration of the microbial community may disrupt the symbiotic relationship and cause or contribute to disease/dysfunction. This new type of microbial diseases is caused by the community as a whole, even though no individual community member(s) can be categorized as a classic pathogen.[1] Alteration of the 'normal' microbiome (dysbiosis) has recently been associated with a variety of high impact diseases including inflammatory diseases (periodontal disease, esophagitis, idiopathic inflammatory diseases, psoriasis), metabolic syndrome-related diseases (diabetes, obesity), immunological disorders (rheumatoid arthritis, food allergy, asthma), cardiovascular disease and cancers.[1–3] Normalization of the altered microbiome by probiotics, antibiotics, prebiotics or microbiome transplantation has been used to evaluate the etiology of dysbiosis and could become a new way to treat these diseases. The approach to identify and characterize the cause differs drastically from traditional approaches aimed at single pathogens.[2]

The first evidence of microbes was presented by Antone van Leeuwenhoek in 1676 using his newly invented microscope. For hundreds of years, observation of bacterial populations did not advance much past visual and culture methods. Culture-dependent methods of microbial identification were the gold standard for bacterial sampling of environments and anatomical sites, and even enabled the formation of Koch's postulate of microbial pathogenesis. But these culture dependent methods are extremely biased towards microbes that can readily be cultured in laboratory settings. The vast majority of microbes are fastidious, and therefore, excluded from any culture-dependent analysis. Even though the science of culturing fastidious microbes has advanced tremendously, culture-dependent methods are, for most complex microbiomes, drastically insufficient for defining microbiome populations.

Cultivation-independent techniques revealed a microbial population far more diverse than previously known. Without the need for culturing before identification, thousands of species of bacteria were revealed in microbiomes previously thought to be dominated by a few cultivable species. The most commonly used cultivation technique was cloning of 16S ribosomal RNA (rRNA) genes of a mixed microbial population into a suitable vector, followed by transformation of bacteria, plating and colony picking, ultimately ending in plasmid purification and Sanger sequencing of the isolated 16S gene. While this research technique was very transformative, it was also limited in several aspects regarding microbiome characterization and very labor intensive. Each individual colony represents one 16S rRNA gene, limiting most studies to relatively few sequences per sample (sometimes in the tens or hundreds). While this gave a larger view of the total microbiome than before, full bacterial diversity could not be explored, especially for rare biosphere bacteria.

The field of microbial ecology has been transformed since the introduction of NGS. Massively high-throughput pyrosequencing has enabled acquisition of millions of sequences from days worth of lab work, the same amount of data that would take years to obtain using cloning methods, and at a fraction of the cost. NGS platforms such as 454 and Illumina make it possible to directly sequence from a pool of amplified 16S rRNA genes. This has been of great use in sequencing 16S rRNA genes from mixed microbial populations, although a drawback to using these methods is the shorter read lengths obtained compared to Sanger sequencing. 454 sequencing has been the preferred platform for 16S surveys, since it is capable of obtaining reads of 450 bases, and soon 1000 bases (compared to 100 base reads for Illumina). A typical 454 FLX run can net 1 000 000 sequences of 450 bases, and samples can be multiplexed using nucleotide barcodes. This, along with dividing the 454 microtiter plate, can allow for sequencing of thousands of samples with coverage of thousands of sequences per

[1]Department of Ecology & Evolutionary Biology, Rice University, Houston, TX 77005, USA; [2]Department of Laboratory Medicine, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA and [3]Departments of Pathology and Medicine, New York University School of Medicine, New York, NY 10016, USA
Correspondence: ZH Pei
E-mail: zhiheng.pei@med.nyu.edu

sample, all for the cost of under $20 000 and 1 day to perform the run. The high-throughput nature of NGS technology has drastically reduced the cost of pricing. An effective way to compare would be to look at price per sequence obtained. Cloning/Sanger sequencing methods on average generate one sequence for a cost of $5 (taking into account not only sequencing, but bacterial culture, plasmid prep, etc.), while a typical 454 run generates one sequence at a cost of about $0.01, a difference of 500-fold.

While the 16S rRNA gene is widely accepted as a biological fingerprint for bacterial species, there are some limitations. Several bacterial species have multiple copies of 16S rRNA genes, which may lead to them being artificially overrepresented in the resultant data.[4] Another limitation is that for some species, 16S rRNA genes might not be the best differentiating gene to choose, for instance, some Anthrax species, have identical 16S rRNA genes, but differ in their extrachromosomal gene content. Focusing on one gene alone also excludes other members of the microbial population, such as archaea, fungi and protists whose small subunit rRNA genes differ significantly from bacterial 16S rRNA genes. Some technical considerations arise from the practice of polymerase chain reaction amplifying 16S rRNA genes for microbiome characterization. These must be taken into account when analyzing data from sequencing runs. There is the possibility of polymerase chain reaction bias during the amplification step which may favor the 16S rRNA genes of some microbes over others, skewing the population structure. This may arise either by preferential annealing of primers to some 16S genes over others, or by more efficient amplification of some 16S rRNA genes over others due to their sequence content (i.e. GC content, secondary structure formation, etc.).[5] Bias due to primer annealing can be overcome by tailoring primers to highly conserved regions, especially when the population of interest is known.[6] In regard to primer selection, it is also important to choose which region of the 16S rRNA gene will be sequenced, since length limitations of NGS prevent sequencing the whole gene and classification accuracy varies with different variable regions of the 16S rRNA gene.[6] Polymerase error is an important factor both in the amplification process and the sequencing process. Taq polymerase error can be minimized during the amplification process by utilizing only high fidelity Taq and keeping the number of cycles relatively low. Errors during sequencing, while a very small proportion may be due to polymerase error, are mostly due to homopolymer errors, which is intrinsic to pyrosequencing. An expected error rate of 1% is the norm for 454 sequencing, which greatly outweighs polymerase errors. Other sequence artifacts may arise from formation of heteroduplex molecules and formation of chimeras. Solutions to these error problems include clustering sequences by 99% sequence similarity (rather than the previously accepted 97%) and employing modified amplification protocols.[5]

While 16S surveys have been useful, the even higher throughput of today's next generation machines has made whole genome shotgun sequencing a more attractive, and available, alternative. This method allows for analysis not only of 16S rRNA genes, but the entire gene content of the microbial population. Using this type of sequence data, it is possible not only to characterize the microbiome by which species are present, but which genes and functional pathways are present. In some cases, that may be even more important data than just which species are present. Additionally, the data from whole genome shotgun sequencing present not only bacterial sequences, but also those of archaea, fungi, protists, some viruses and eukaryotes—providing a true microbial population and metagenome instead of just a bacterial population.

It should also be noted that microbiome sequencing results between people, and even from different samples from the same individual will usually differ, and it is not unusual for results to not be 100% replicable. A certain portion, sometimes very small, of the microbiome is in flux due to temporal or environmental factors, and thus, there is always some amount of variability involved when sampling the microbiome. The important consideration is that the core components remain somewhat constant and the relative abundance range of more abundant species remains within range and does not fluctuate so much as to give different interpretations of results.

Although getting the whole genome shotgun sequences may be readily done by most researchers, the data analysis is not trivial. Not only is there a large volume of data, which presents its own computational challenges, but many sequences obtained will have no representative within the database, making their usefulness limited. Also, assembly of genes and genomes from this data is complicated by the short read lengths and repetitive DNA elements in many genomes, a problem that has plagued genome assembly teams, especially when the sequence reads are shorter than the repeats. To obtain complete circular genomes could require deeper sequencing, including sequencing across gaps or large insert plasmid libraries.

However, the ultimate solution relies on the development of new technical platforms that can generate long reads. The accuracy of microbiome assessment can be significantly improved by eliminating amplification bias and increasing read length. There are several new platforms in the development in this field. Some developing sequencing platforms are miniaturized to fit in the palm of your hand and plug into a portable computer. These devices use nanopores to sequence individual strands of DNA, without amplification, at competitive accuracies. Another new platform sequences single DNA molecules in real time with no need for polymerase chain reaction amplification which produces read lengths greater than 3000 base pairs and allows users to crack down the types of complex, repetitive sequence that perplex the short-read sequencing platforms. Microfluidic based devices have also been able to provide genome analysis of microbial consortia from diverse environmental samples including a marine enrichment culture, deep-sea sediments and the human oral cavity.[7]

In the short span of about 5 years, NGS has evolved from a promise, to a reality, to a necessity. It has almost reached a point where reviewers expect NGS for microbial analysis rather than antiquated methods whose results are superficial and lacking. It seems that, until some other innovative technology supplants it, NGS-based genome analysis is a must for deep microbial population characterization.

ACKNOWLEDGMENTS

The work was supported by grants UH3CA140233, R01CA159036, R01AI063477 and U19DE018385 from the National Cancer Institute and the National Institute for Allergy and Infectious Diseases, and National Institute of Dental and Craniofacial Research.

1 Yang L, Lu X, Nossa CW, Francois F, Peek RM, Pei Z. Inflammation and intestinal metaplasia of the distal esophagus are associated with alterations in the microbiome. *Gastroenterology* 2009; **137**: 588–597.
2 NIH HMP Working Group, Peterson J, Garges S *et al*. The NIH Human Microbiome Project. *Genome Res* 2009; **19**: 2317–2323.
3 Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. *Nat Rev Genet* 2012; **13**: 260–270.
4 Pei AY, Oberdorf WE, Nossa CW *et al*. Diversity of 16S rRNA genes within individual prokaryotic genomes. *Appl Environ Microbiol* 2010; **76**: 3886–3897.
5 Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF. PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Appl Environ Microbiol* 2005; **71**: 8966–8969.

6 Nossa CW, Oberdorf WE, Yang L *et al*. Design of 16S rRNA gene primers for 454 pyrosequencing of the human foregut microbiome. *World J Gatroenterol* 2010; **16**: 4135–4144.
7 Leung K, Zahn H, Leaver T *et al*. A programmable droplet-based microfluidic device applied to multiparameter analysis of single microbes and microbial communities. *Proc Natl Acad Sci USA* 2012; **109**: 7665–7670.